Deutsche Bahn

The not-so-efficient train network.

Iron Hack Final Project - Ray Pham

About Deutsche Bahn

Deutsche Bahn is the national railway company of Germany.

It is the largest railway operator in Europe and the second-largest in the world by revenue.

Deutsche Bahn is a major employer in Germany with over 324,000 employees.

Generates over €56B in revenue per year.

Quick Facts

It operates a network of over 33,000 kilometres of track. This network includes over 3,000 stations and serves over 2,000 cities and towns.

It operates more than 40,000 trains daily. Tickets for the long-distance service range between 29€ to 150€ per journey.

Deutsche Bahn carries almost 2 billion passengers per year.

German Not-so-efficient

Deutsche Bahn once had a reputation akin to the entire image of Germany, punctuality and efficiency.

However in the past years, due to the aging and neglected infrastructure, this has become a problem for train reliancy and punctuality.

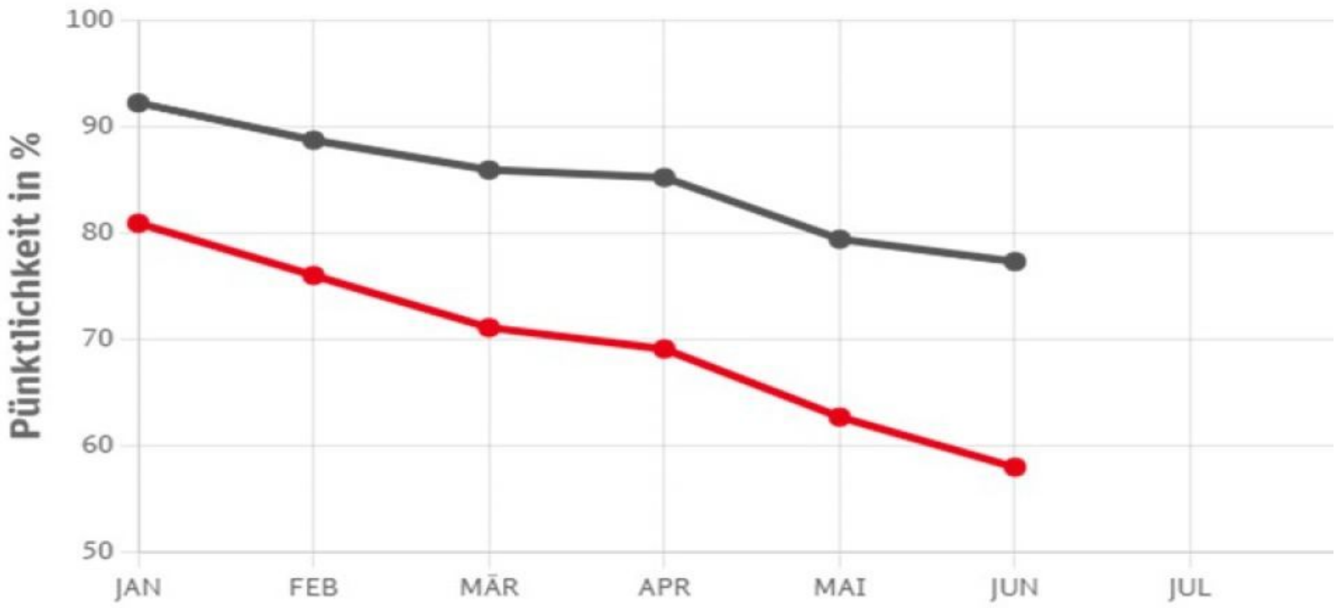
80,6%
long-distance trains on time in 2022.

58%
long-distance trains on time in 2023.

In Germany, a train is considered ‘on-time’ if it arrives no later than 6 minutes of it’s scheduled time.

Source: https://www.railtech.com/all/2023/12/05/deutsche-bahn-punctuality-drops-to-record-low/?gdpr=accept

2023 January to June



Objectives

This analytic project is to understand the efficiency of the Deutsche Bahn train network on key routes, by analysing key dimensions such as delays, distance, timings.

1. Which stations have the most delays
2. If time of travel causes delays
3. Does distance of between affect delays
4. Build a model that can predict potential delays

Unfortunately I wasn’t able to retrieve the passenger information and due to the amount of travels and stations, I focused the project on key cities.

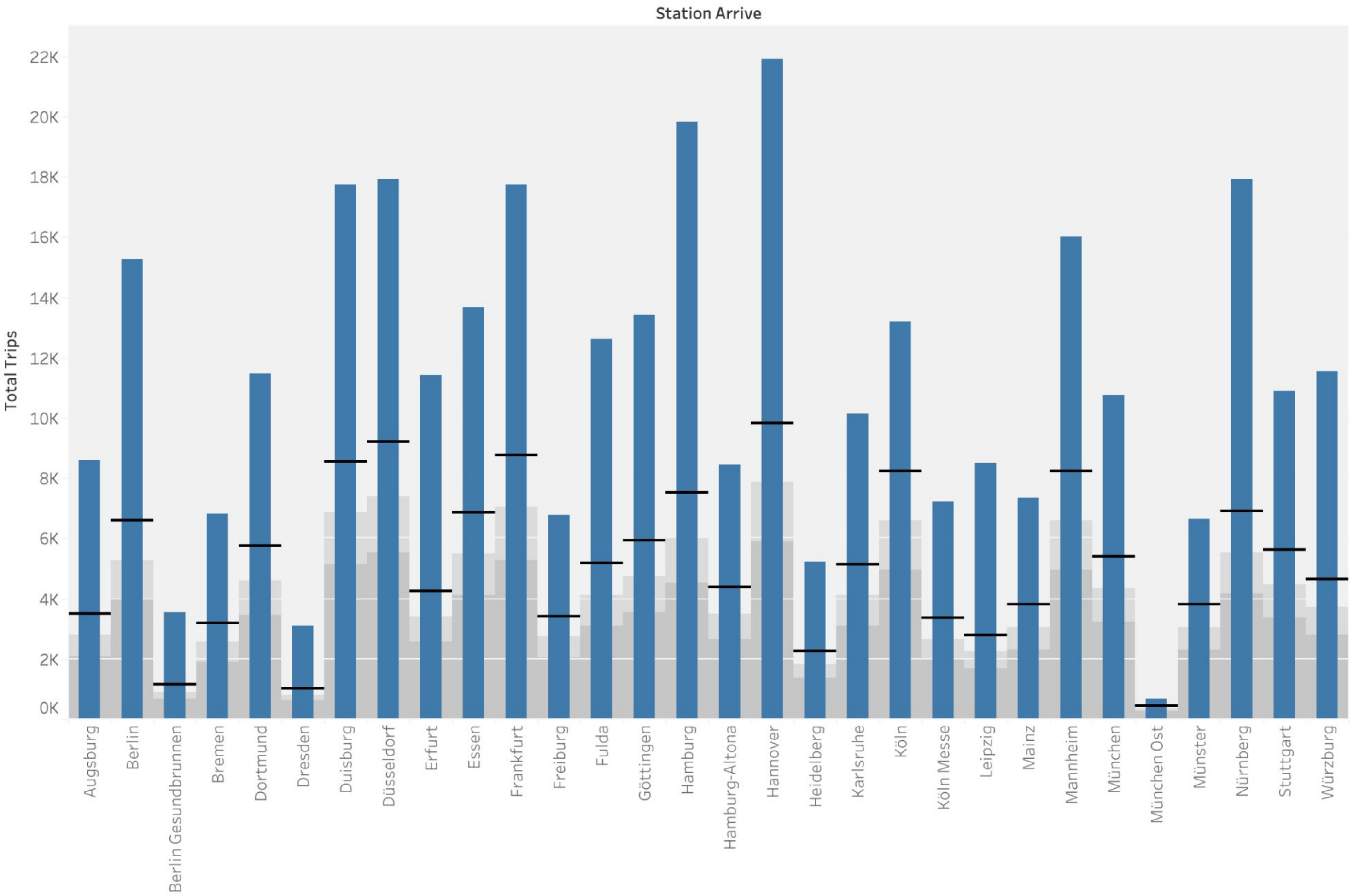
The data observed

Period: May - August 2022
Trips: 428595 (Time/Date)
Routes: 704 (Depart/Arrive Station)
Stations: 36 (Germany and EU)
Distance Calculated (Lat/Lon)

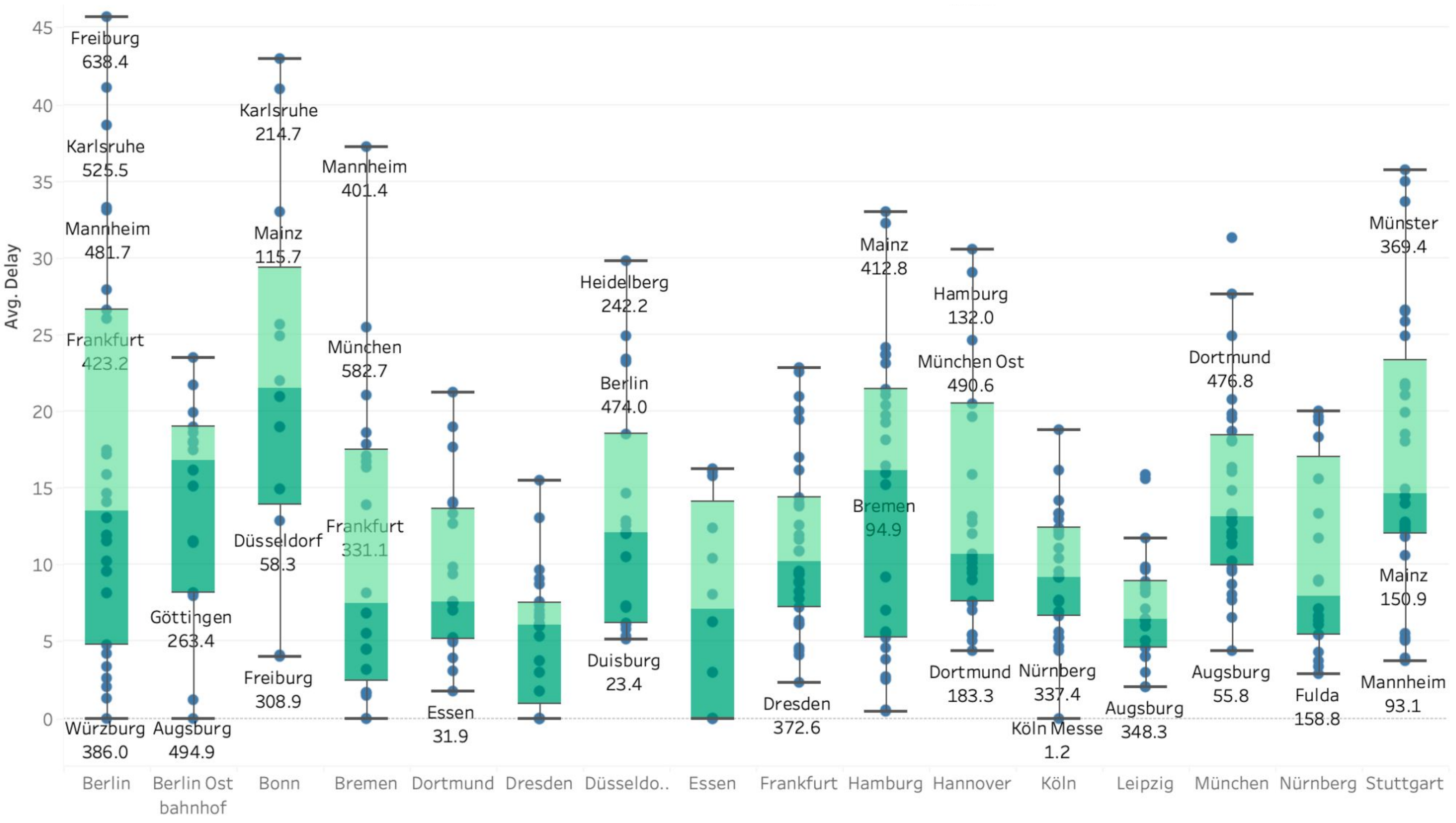
Train Types:
ICE = InterCity Express (DE)
IC = InterCity (DE)
EC = EuroCity (EU)
FLX = FlixTrain (EU)
ECE = EuroCity Express (DE)
EN = EuroNight (EU)

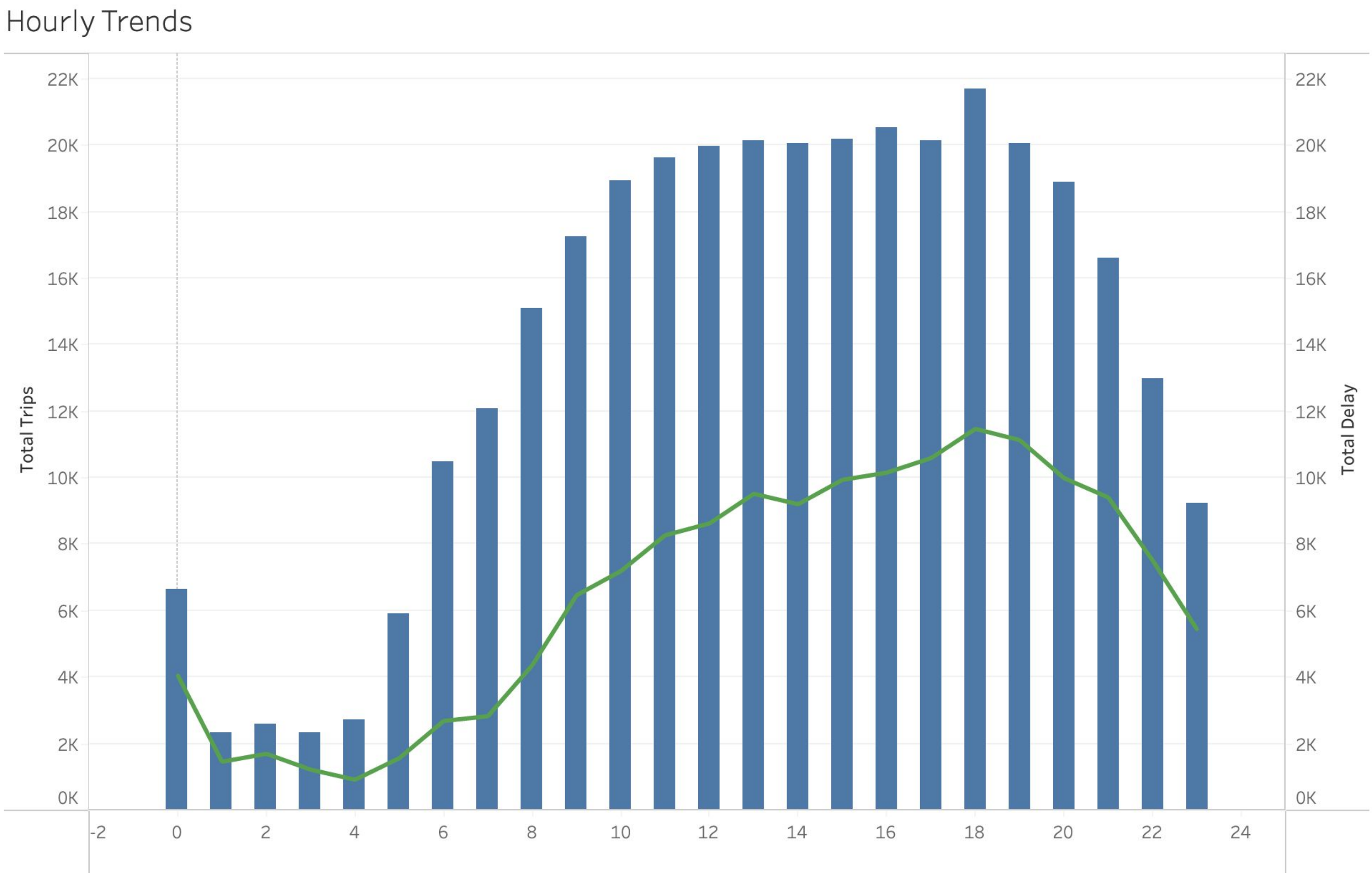
Top 10 delay routes

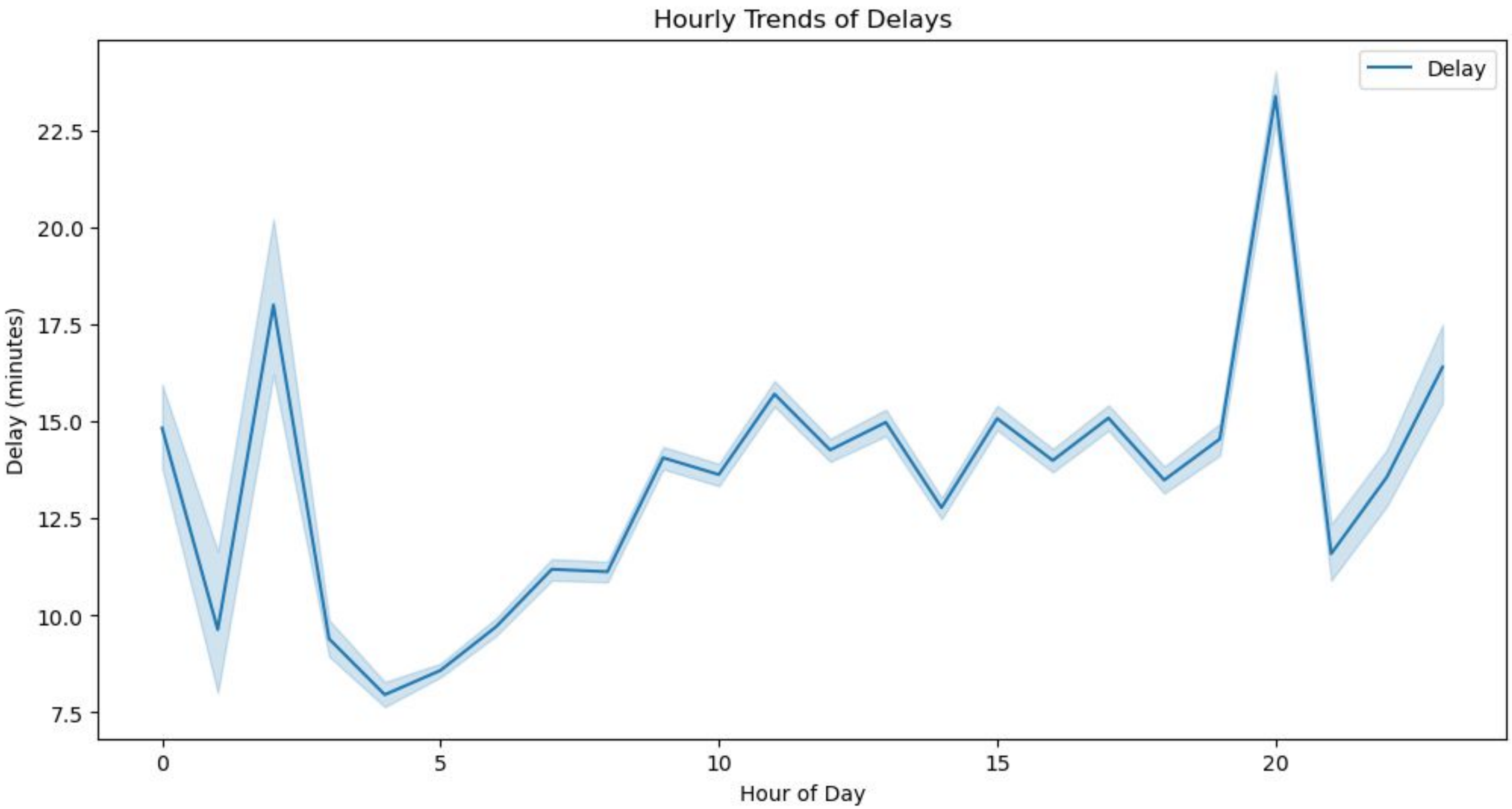
Arrival	Departure	Trips	Delays	Delay %	Avg. Min	Distance KM
Hamburg	München	4826	2649	54.89	28.22	611.578
Nürnberg	München	7357	2321	31.55	18.82	149.285
Hamburg-Altona	München	4204	2263	53.83	26.81	612.556
München	Hamburg-Altona	4268	2237	52.41	29.17	612.556
Mannheim	Hamburg-Altona	2921	2212	75.73	31.8	464.138
Hannover	Köln	3394	2161	63.67	24.67	249.447
Nürnberg	Hamburg-Altona	4198	2037	48.52	25.87	463.588
Hannover	Hamburg-Altona	4116	1983	48.18	19.77	131.495
Düsseldorf	München	2669	1965	73.62	25.93	484.48'
Frankfurt	München	3512	1952	55.58	21.99	303.663


Station Arrive
Total Trips
22K
20K
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
Augsburg
Berlin
Berlin Gesundbrunnen
Bremen
Dortmund
Dresden
Duisburg
Düsseldorf
Erfurt
Essen
Frankfurt
Freiburg
Fulda
Göttingen
Hamburg
Hamburg-Altona
Hannover
Heidelberg
Karlsruhe
Köln
Köln Messe
Leipzig
Mainz
Mannheim
München
München Ost
Münster
Nürnberg
Stuttgart
Würzburg

Delays between German cities




Hourly Trends
Total Trips
Total Delay
22K
20K
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
-2
0
2
4
6
8
10
12
14
16
18
20
22
24

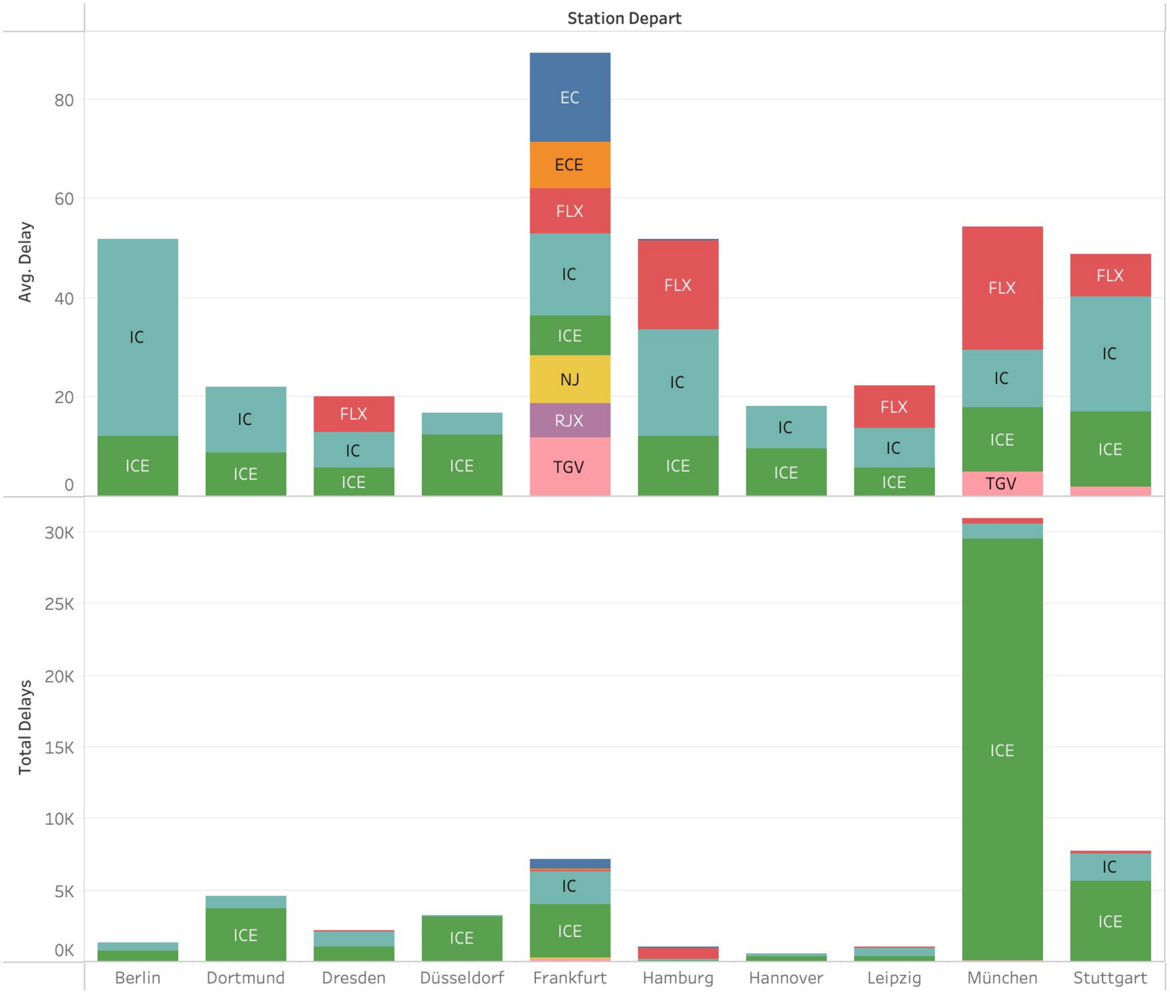


Hourly Trends of Delays
Delay
22.5
20.0
17.5
15.0
12.5
10.0
7.5
Delay (minutes)
0
5
10
15
20
Hour of Day

Train Types

Observing the types of trains with the most delays, we can see that the ICE service from Munich contributes to the majority of delays.

But on average, the delay time is roughly equal across all station departures.



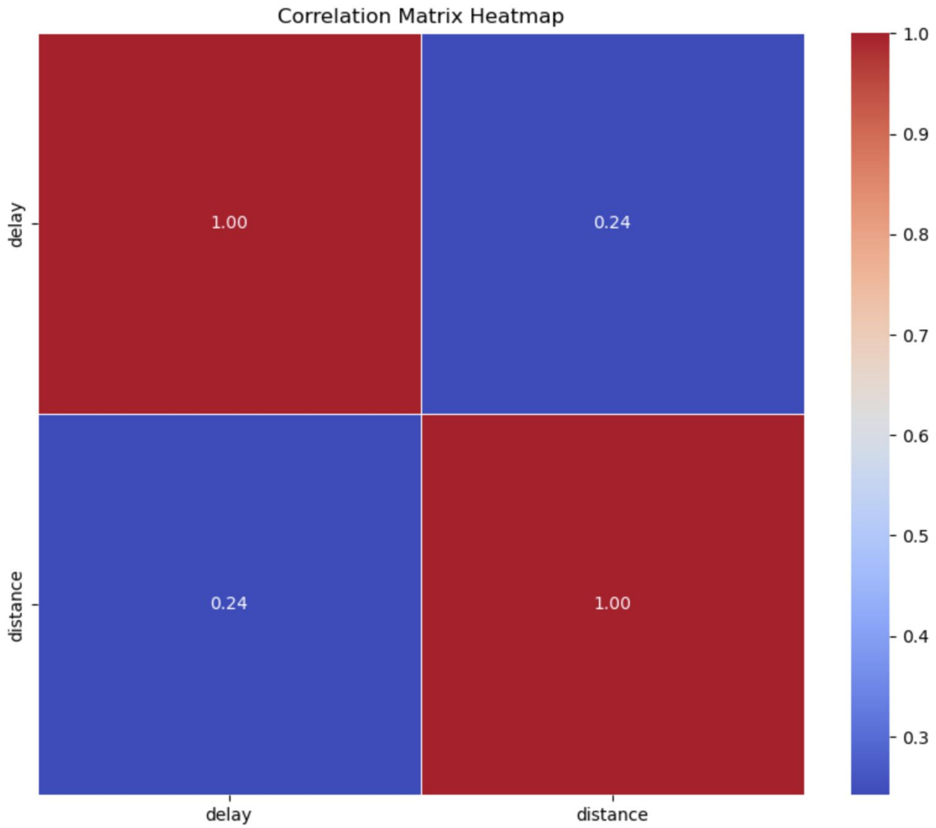
Station Depart	Train Type	Trips	Avg. Delay	Total Delays	Canceled
Berlin	IC	24,286	40	532	6
	ICE	22,698	12	760	143
Dortmund	IC	23,609	13	875	150
	ICE	92,978	9	3,703	402
Dresden	FLX	2,064	7	116	6
	IC	23,933	7	1,013	164
	ICE	22,588	6	1,091	201
Düsseldorf	IC	58	4	5	3
	ICE	85,188	12	3,228	532
Frankfurt	EC	17,486	18	657	84
	ECE	3,437	9	151	72
	FLX	996	9	22	65
	IC	63,821	16	2,311	242
	ICE	89,839	8	3,730	1,033
	NJ	648	10	26	2
	RJX	1,117	7	54	11
	TGV	5,365	12	240	13
Hamburg	EC	7	0	0	0
	FLX	35,227	18	878	127
	IC	2,006	21	50	1
	ICE	3,835	12	141	6
Hannover	IC	6,042	9	194	13
	ICE	9,308	10	389	62
Leipzig	FLX	2,951	8	94	12
	IC	13,749	8	561	86
	ICE	7,366	6	368	60
München	FLX	14,388	25	337	39
	IC	23,750	12	1,035	238
	ICE	825,258	13	29,475	2,734
	TGV	1,582	5	72	8
Stuttgart	FLX	8,184	9	249	30
	IC	67,202	23	1,861	179
	ICE	170,189	15	5,636	1,033
	TGV	13	2	1	0

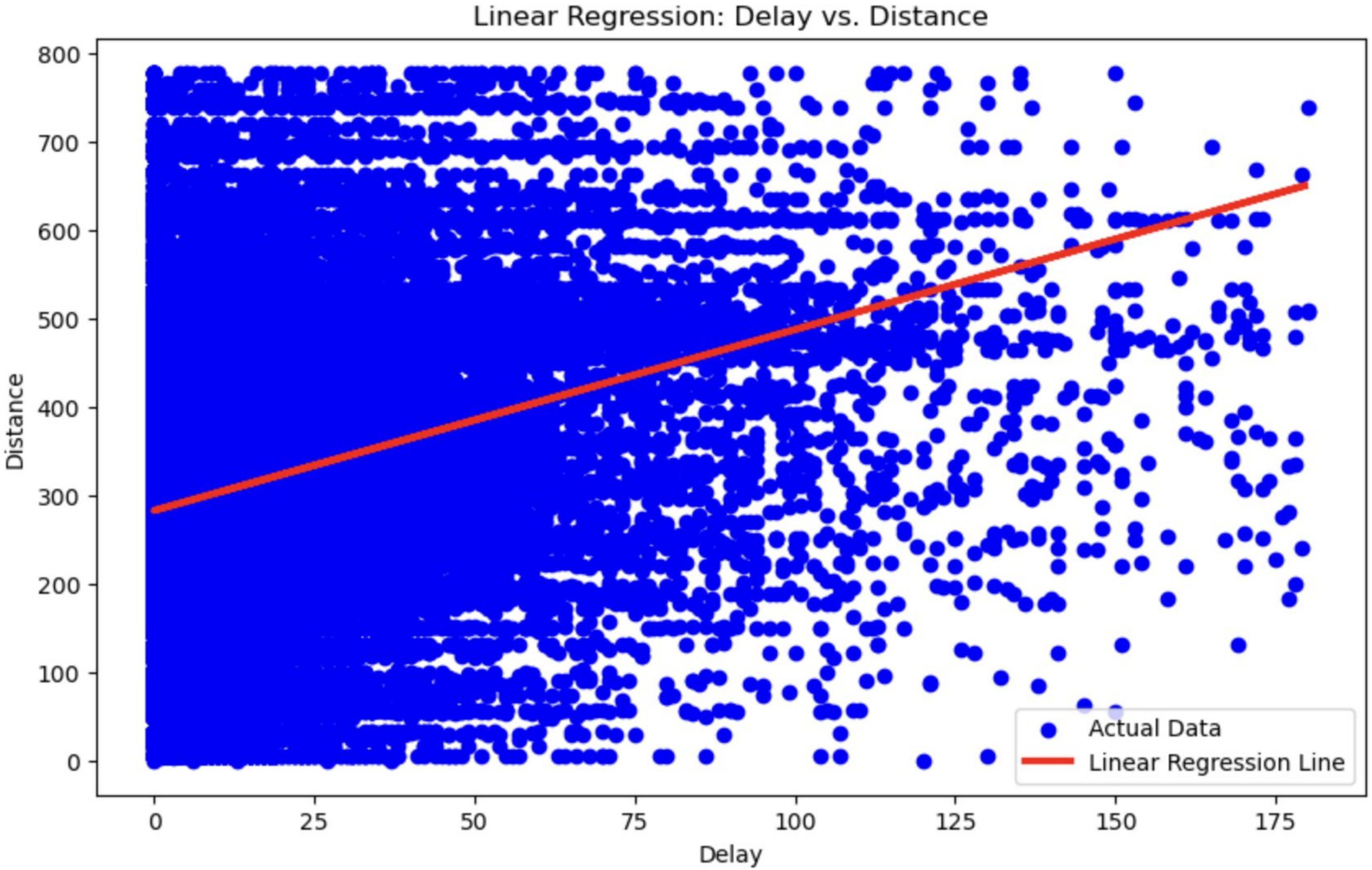
Model Testing

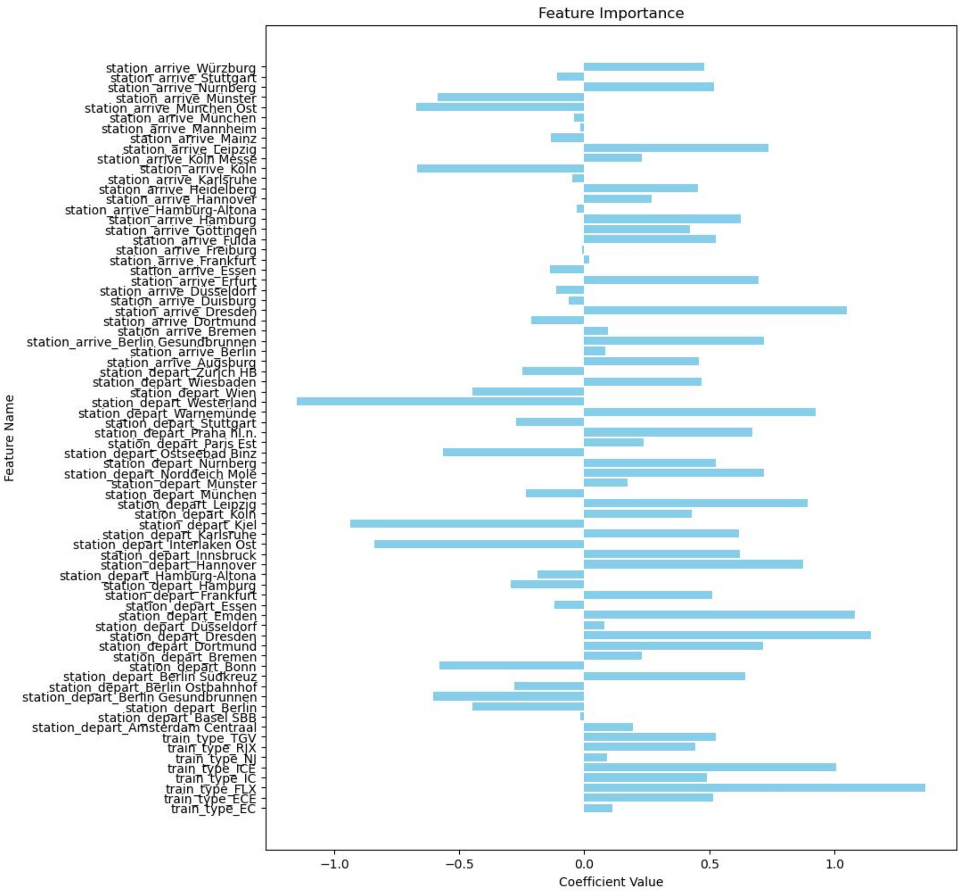
Looking at the initial analysis, I can already observe that there isn't a very clear pattern between features and numericals, so tested a few different models to see, if based on one numerical value (distance) could be enough to build a model.

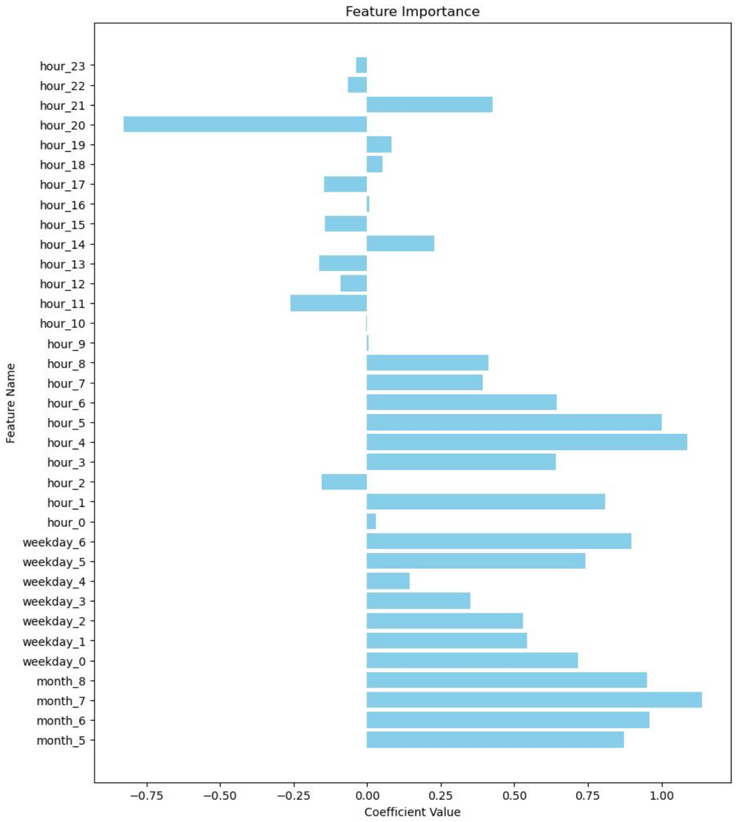
For the test for the features was used

- Station (Destination & Arrival)
- Day of Week
- Hour of Day
- Type of Train (ICE, IC etc)


Correlation Matrix Heatmap
delay
distance
1.00
0.24
0.24
1.00
delay
distance
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3


Linear Regression: Delay vs. Distance
Distance
Delay
0
100
200
300
400
500
600
700
800
0
25
50
75
100
125
150
175
Actual Data
Linear Regression Line


Feature Importance
station_arrive_Würzburg
station_arrive_Stuttgart
station_arrive_Nürnberg
station_arrive_Münster
station_arrive_München Ost
station_arrive_München
station_arrive_Mannheim
station_arrive_Mainz
station_arrive_Leipzig
station_arrive_Köln Messe
station_arrive_Köln
station_arrive_Karlsruhe
station_arrive_Heidelberg
station_arrive_Hannover
station_arrive_Hamburg-Altona
station_arrive_Hamburg
station_arrive_Göttingen
station_arrive_Fulda
station_arrive_Freiburg
station_arrive_Frankfurt
station_arrive_Essen
station_arrive_Erfurt
station_arrive_Düsseldorf
station_arrive_Duisburg
station_arrive_Dresden
station_arrive_Dortmund
station_arrive_Bremen
station_arrive_Berlin Gesundbrunnen
station_arrive_Berlin
station_arrive_Augsburg
station_depart_Zürich HB
station_depart_Wiesbaden
station_depart_Wien
station_depart_Westerland
station_depart_Warnemünde
station_depart_Stuttgart
station_depart_Praha hl.n.
station_depart_Paris Est
station_depart_Ostseebad Binz
station_depart_Nürnberg
station_depart_Norddeich Mole
station_depart_Münster
station_depart_München
station_depart_Leipzig
station_depart_Köln
station_depart_Kiel
station_depart_Karlsruhe
station_depart_Interlaken Ost
station_depart_Innsbruck
station_depart_Hannover
station_depart_Hamburg-Altona
station_depart_Hamburg
station_depart_Frankfurt
station_depart_Essen
station_depart_Emden
station_depart_Düsseldorf
station_depart_Dresden
station_depart_Dortmund
station_depart_Bremen
station_depart_Bonn
station_depart_Berlin Südkreuz
station_depart_Berlin Ostbahnhof
station_depart_Berlin Gesundbrunnen
station_depart_Berlin
station_depart_Basel SBB
station_depart_Amsterdam Centraal
train_type_TGV
train_type_RJX
train_type_NJ
train_type_ICE
train_type_IC
train_type_FLX
train_type_ECE
train_type_EC
Feature Name
−1.0
−0.5
0.0
0.5
1.0
Coefficient Value


Feature Importance
hour_23
hour_22
hour_21
hour_20
hour_19
hour_18
hour_17
hour_16
hour_15
hour_14
hour_13
hour_12
hour_11
hour_10
hour_9
hour_8
hour_7
hour_6
hour_5
hour_4
hour_3
hour_2
hour_1
hour_0
weekday_6
weekday_5
weekday_4
weekday_3
weekday_2
weekday_1
weekday_0
month_8
month_7
month_6
month_5
Feature Name
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Coefficient Value

Model Testing

Logistic
MSE:
372.4184

Random Forest
MSE:
464.4855

Decision Tree
MSE:
525.8380

Really bad scores on the model testing.

Testing a Classification Model

	precision	recall	f1-score
0	0.71	0.72	0.71
1	0.67	0.65	0.66

To approach a different model,
0 = No, 1 = Yes binary classification was added if the the percentage if the train will be delayed for longer than 6 minutes.

First Learnings

My first learnings is with the dataset, and after the wrangling i performed, it didn’t do enough to build a model that was able to predict delays on an accurate basis.

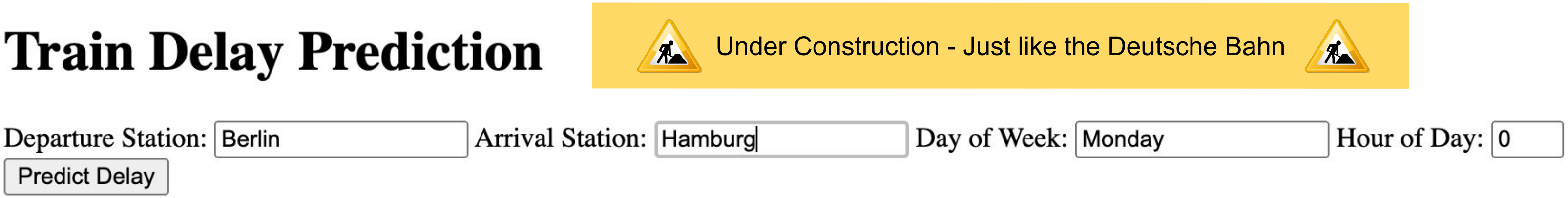
Things I could try to apply as next steps:

- Enrich the dataset with more variables (such as passenger count per station per day)
- Test different approaches with dates (clustering between weekday and weekend).
- Add an additional numerical dimension by calculating the total number of trips between key certains.

Conclusion

My assumption of distance impacts delays applies in some cases only. Time of travel has a minor effect on the likelihood that a train will be late (early morning as opposed to rush hour which also affects the average delay duration) but the impact follows the trip frequency trend.

More time and data is needed, as well as finding a way to cluster the data that would allow myself to build a better prediction model.



Danke.